The reinforcement heuristic in normal form games

Alós-Ferrer, Carlos ; Ritschel, Alexander

Abstract: We analyze simple reinforcement-based behavioral rules in 3×3 games through choice data and response times. We argue that there is a large overlap between reinforcement-based heuristics (win-stay, lose-shift) and the more "rational" behavioral rule of myopic best reply. However, evidence from response times shows that choices in agreement with the common prescription of those rules are comparatively fast, and choices of the form "lose-shift" occur more frequently for larger differences with bygone payoffs. Both observations speak in favor of reinforcement processes as a cognitive shortcut for apparent myopic best reply, and advise caution when interpreting behavioral results in favor of optimizing behavior.

DOI: https://doi.org/10.1016/j.jebo.2018.06.014

The Reinforcement Heuristic in Normal Form Games*

Carlos Alós-Ferrer† University of Zurich

Alexander Ritschel‡ University of Zurich

This Version: June 2018

Abstract

We analyze simple reinforcement-based behavioral rules in 3×3 games through choice data and response times. We argue that there is a large overlap between reinforcement-based heuristics (win-stay, lose-shift) and the more "rational" behavioral rule of myopic best reply. However, evidence from response times shows that choices in agreement with the common prescription of those rules are comparatively fast, and choices of the form "lose-shift" occur more frequently for larger differences with bygone payoffs. Both observations speak in favor of reinforcement processes as a cognitive shortcut for apparent myopic best reply, and advise caution when interpreting behavioral results in favor of optimizing behavior.

JEL Classification: C72 · C91

Keywords: Reinforcement · Myopic Best Reply · Response Times · Decision Processes

*The authors thank Arno Riedl, John Smith, two anonymous referees, and conference participants at the THEEM 2017 in Kreuzlingen and the NPE conference 2017 in Antwerpen for helpful comments and suggestions. The authors also gratefully acknowledge financial support from the German Research Foundation (DFG) under project Al-1169/4, part of the Research Unit "Psychoeconomics" (FOR 1882). The Department of Economics at the University of Cologne gratefully acknowledges financial support from the DFG to build the Cologne Laboratory for Economic Research.

†Corresponding author: carlos.alos-ferrer@econ.uzh.ch. Department of Economics, University of Zurich. Blümlisalpstrasse 10, 8006 Zurich, Switzerland.

‡Department of Economics, University of Zurich. Blümlisalpstrasse 10, 8006 Zurich, Switzerland.

1 Introduction

Reinforcement is one of the most basic processes underlying human learning. Accordingly, it has received widespread attention in psychology, going back to Thorndike's (1911) "law of effect," neuroscience (e.g. Holroyd and Coles, 2002; Schönberg et al., 2007), and computer science (Sutton and Barto, 1998). Within microeconomics and game theory, it has been frequently studied as a boundedly-rational behavioral rule (see, e.g. Börgers and Sarin, 1997; Erev and Roth, 1998), as have been other rules, e.g. imitation or myopic best reply (Weibull, 1995; Fudenberg and Levine, 1998). The simplest version of reinforcement learning can be viewed as a heuristic which takes past experiences into account for the choice of upcoming actions and prescribes a shift from actions linked to negative experiences to actions associated with positive rewards: that is, "win-stay, lose-shift." This heuristic induces a bias towards past-high-reward actions which can conflict with rational behavior (outcome bias; Baron and Hershey, 1988; Dillon and Tinsley, 2008).

Evidence from neuroscience shows that reinforcement-based decisions occur extremely fast in the human brain (Schultz, 1998; Holroyd and Coles, 2002). Indeed, reinforcement is a textbook example of an *automatic process*, as conceived in dual-process theories from psychology (see, e.g., Kahneman, 2003; Strack and Deutsch, 2004; Alós-Ferrer and Strack, 2014). Those theories define automatic (or intuitive) processes as immediate, fast, unconscious, and efficient in the sense of requiring few cognitive resources. For instance, these processes capture impulsive reactions and behavior along the lines of stimulus-response schemes. The dual-process approach postulates that human decisions are mainly influenced by automatic processes and so-called *controlled* (or deliberative) processes. The latter are seen as slow, consuming cognitive resources, not instigated immediately, and reflected upon consciously. Explicit maximization of expected rewards, if conceptualized as a process, would exhibit many if not all of those characteristics.

The relevance of reinforcement for economic decision making was illustrated by Charness and Levin (2005) in a binary-choice, belief-updating task where mistakes (deviations from optimization under correct Bayesian updating) could be traced to a reinforcement heuristic. In essentially the same paradigm, Achtziger and Alós-Ferrer (2014) found evidence of the conflict between reinforcement and rational optimization in the form of response time asymmetries as predicted by an explicit dual-process model. Recent psychophysiological work (Achtziger et al., 2015) found direct evidence of neural correlates of reinforcement in this paradigm and studied their relation to economic incentives. Further studies relying on this paradigm have examined the interaction of reinforcement and decision inertia (Alós-Ferrer et al., 2016), the influence of framing on reinforcement decisions (Alós-Ferrer et al., 2017), and the regulation of reinforcement processes through motivational interventions (Hügelschäfer and Achtziger, 2017).

In this work, we take a further step in the study of reinforcement heuristics in economic settings by moving beyond binary-choice tasks and studying the explicit relation

between reinforcement and myopic payoff maximization in strategic decisions. Hence, we study reinforcement processes in a more complex setting which results in longer decisions times than, e.g., standard neuropsychological experiments. We concentrate on two-player, 3×3 asymmetric normal form games. In this setting, the microeconomics literature has devoted a great deal of attention to myopic best reply, a behavioral rule which maximizes the own payoff assuming the other player will repeat her action, and which can be assumed to have a more deliberative/controlled nature than reinforcement.

Previous work has analyzed paradigms where, by design, reinforcement and more deliberative behavior could either conflict or be aligned (e.g. Achtziger and Alós-Ferrer, 2014). In other settings, however, there might be a great degree of overlap between the prescriptions of reinforcement and those of myopic best reply. In the present work we specifically explore to what extent reinforcement can act as a shortcut for (apparent) optimization in strategic situations with explicit feedback. Suppose a player's last action delivered the best possible payoff. Reinforcement will then prescribe to repeat the choice (win-stay). By definition, however, that choice is the best reply if the opponent stays put. Likewise, suppose the last action did not deliver the best possible payoff. Reinforcement will prescribe to choose a different action (lose-shift). But, again by definition, the current choice cannot be the myopic optimum, and hence myopic best reply also prescribes to shift. In principle, the "shift" prescribed by reinforcement is arbitrary, but if payoffs are observable (as, e.g., if the payoff table is known), the observable maximum becomes salient and the shift will often be in its direction, leading to an apparent myopic best reply. In view of these observations, we postulate that reinforcement processes might often act as cognitive shortcuts resulting in choices indistinguishable from myopic best reply.[1]

If choices coming from reinforcement and those prescribed by myopic best reply cannot be distinguished, how can this hypothesis be substantiated? There are two possible avenues. The first relies on response times. As explained above, reinforcement processes are automatic and can be expected to lead to shorter response times than alternative processes. In contrast, myopic best reply involves explicit maximization and can be assumed to be deliberative, hence relatively slow. Hence, if the choices favored by both processes are actually due to the involvement of reinforcement processes, one should expect shorter response times (compared to other choices), while if they are due to explicit maximization, response times should be longer.

The second avenue is bygone payoffs. Suppose a player obtains a payoff which is not the maximum possible one given the opponent's strategy. If that maximum payoff is observable, the deviation with respect to it is a cardinal measure of experienced disappointment. Define experienced regret as the difference between the maximum possible payoff (that of the best reply) and the actually obtained payoff. By definition, regret is zero if and only if the player has chosen a best reply, and strictly positive if not (this will

[1]Indeed, the obvious evolutionary reason for the existence of automatic processes is that in certain situations they are adapted and support (near-)optimal decisions while saving cognitive costs.

be directly observable in our experiment). Myopic best reply, considered as a behavioral rule, prescribes to change strategy whenever a best reply has not been chosen. In contrast, reinforcement processes are stimulus-response mappings which take the win-loss information as an input. The loss information also carries a measurement of stimulus strength which in turn yields a variation in the responses. Hence, standard formalizations of reinforcement take the probability of a shift to be increasing in the degree of the loss, which is just the experienced regret as defined above. That is, reinforcement should be triggered more often for larger experienced regret. Hence, if observed choices follow from reinforcement processes, one should observe a dependence on experienced regret.

To study these questions, we conducted an experiment ($N = 144$) where participants played 3×3 games against other players repeatedly. In order to isolate the decision processes of interest, in our experiment players had full knowledge of their own payoff tables, but were not aware of the payoffs of the opponents. In this way, we aimed to eliminate a number of potential confounds, as e.g. imitation or social preferences. Also, in this simple design the maximum (bygone) payoff associated with a choice is directly observable, and regret is simply the difference between the highest payoff associated with the opponent's choice and the actually received payoff. To make this even simpler, each of the different payoff tables used in the experiment contain only three different numerical payoffs, hence maximum payoffs and regret levels are easily observable.

The experiment was divided in short parts of 13 rounds each, and after each part both the opponent and the payoff table were changed. This was an explicit design decision in order to prevent convergence or long-run effects, and rather concentrate on the decision processes.

Our paper is related to several literature strands. Within the reinforcement learning literature, we focus on simple stimulus-response behavioral rules of thumb, capturing the basic idea of a "win-stay, lose-shift" stimulus-response mapping. Of course, the literature on reinforcement also encompasses more involved models. For instance, the game-theoretic cumulative proportional reinforcement of Laslier et al. (2001) postulates that actions are chosen with probabilities proportional to their cumulative payoffs obtained in the past with that move. Modern reinforcement learning models in computational neuroscience (see Daw, 2012, for an introduction) often include additional factors, e.g. through a perseveration parameter capturing the tendency to repeat or avoid recently chosen actions (e.g., Gershman et al., 2009; Wimmer et al., 2012). Our interest, however, is on Markov behavioral rules conceived as mappings from information at t (e.g., payoffs and actions of all players) to behavior at $t + 1$ (probabilities of the feasible actions), as standard in the literature of learning in games (Fudenberg and Levine, 1998). Within this literature, such behavioral rules include imitation (e.g., Vega-Redondo, 1997; Alós-Ferrer and Weidenholzer, 2008) and myopic best reply (e.g. Kandori and Rob, 1995; Alós-Ferrer and Weidenholzer, 2007). This focus is consequential for the analysis. On the assumption that behavior follows Markov rules, one can concentrate on the estimation of probabilities of actions given only the relevant information, e.g. realized and

bygone payoffs and actions in the last period. For instance, under such maintained hypotheses, the matching within an experiment is irrelevant, as behavior is assumed to rely only on the information presented to the player (we will, however, also control for matching in our analysis).

Our experiment is also related to the literature on multi-armed bandits (Gittins, 1979, 1989), in the sense that players repeatedly choose an action out of three possibilities. However, contrary to this literature we are not interested in optimal (normative) dynamic strategies, but rather on the actual (one-shot) decision processes employed by human beings in our setting. Since players were explicitly told that the opponent was human, our experiment was not framed in terms of intertemporal optimization or uncovering of optimal actions. Indeed, in our data, we see little evidence for intertemporal optimization or learning effects. On the contrary, we argue that, in our context, even what might be interpreted as one-shot optimization might be actually supported by simpler decision processes.

A more-closely related literature has examined "decisions from experience," where subjects make repeated individual decisions in stochastic frameworks but learn the underlying probabilities by making decisions (as opposed to decisions from description, where the priors are induced; Hertwig et al., 2004). Although our setting is interpersonal (subjects play against a human being and not a fixed distribution), insights from this literature are informative. In particular, Erev and Haruvy (2016, Section 1.1.5) remark that decision inertia plays an important role (recall also Alós-Ferrer et al., 2016), which in our setting would lead to higher rates of win-stay choices compared to lose-shift ones. This is indeed found in our data. Erev and Haruvy (2016) also list "surprise-triggers-change" as one of the main reasons for not repeating the last choice; that is, the probability of inertia decreases when recent outcomes are surprising. In our case, a larger experienced regret plays the role of surprise. The mirror image of this effect is "negative recency," which is sometimes observed in decisions from experience and implies higher shift rates after surprising positive outcomes (in our case, no regret), even though those reinforce the last choice.

The remainder of the paper is structured as follows. Section 2 presents the experimental design and procedures. Section 3 presents the results, analyzing both choice data and response times. Section 4 concludes.

2 Experimental Design and Procedures

We conducted 6 sessions with 24 participants each for a total of 144 subjects (91 female) at the Cologne Laboratory for Economic Research (CLER). Participants were recruited via ORSEE (Greiner, 2015) from the student population of the University of Cologne excluding students of psychology. The average age was 22.97 years (median 23, range 18–50). The experiment was programmed in z-Tree (Fischbacher, 2007) and sessions

Game 1

	$\sim$	$\bullet$	x
o	6,2	4,4	2,6
#	2,6	6,2	4,4
$\div$	4,4	2,6	6,2

Regret $\in \{0, 2, 4\}$

Game 2

	$\sim$	$\bullet$	x
o	7,1	4,4	1,7
#	1,7	7,1	4,4
$\div$	4,4	1,7	7,1

Regret $\in \{0, 3, 6\}$

Game 3

	$\sim$	$\bullet$	x
o	6,1	5,5	1,6
#	1,6	6,1	5,5
$\div$	5,5	1,6	6,1

Regret $\in \{0, 1, 5\}$

Table 1: The three games used in the experiment.

lasted on average 60 minutes. The average payoff was 12.28 EUR (SD= 0.94) including a show-up fee of 2.50 EUR.

The experiment involved three different 3×3 normal form games (see Table 1). The games had a cyclical structure and the three strategies were neutrally labeled: $\circ$, #, $\div$ for player 1, and $\sim$, $\bullet$, x for the opponent. Each table uses only three different payoffs, and each outcome is clearly attainable for each action the opponent possibly plays. Hence, maximum bygone payoffs and regret levels (given the actual opponent's strategies) are directly and easily observable.

Each subject played a total of 39 rounds divided into three different parts of 13 rounds each. In each part, a subject faced a fixed game and played against a fixed partner. The subject saw a 3×3 payoff table with only her own payoffs. That is, she was not informed about the payoffs of the other player and hence imitation was not feasible, and social preferences were not a concern. The subject chose one of the three actions ($\circ$, #, $\div$) in each round. Rematching was done within blocks of four players after 13 rounds and a new game, i.e. payoff table, was presented. The payoff tables were always reordered and relabeled in such a way that every player saw herself as player 1. The order of the labels for the own strategies was counterbalanced among subjects. Subjects were paid for each decision in all rounds. Alternatively, we could have paid one randomly selected round; Charness et al. (2016) have shown that relying on one or the other method does not significantly affect behavior.[2]

The games were designed to generate different levels of "regret," defined as the difference between the maximum possible payoff and payoff earned in the previous round. The levels ranged from 0 to 6 (see Table 1). By construction, in our games a best reply always reached the maximum payoff available in the payoff table. This was done on purpose to avoid alternative definitions of experienced regret and, hence, potential alternative rules based on reinforcement heuristics.

After each round, the actual play from the previous round was revealed. Subjects saw their own choice, the opponent's choice, and their own payoff. This information remained on-screen while they made their choice for the next round (obviously, there

[2]However, Azrieli et al. (2018, Appendix C) conclude that in a dynamic setting with feedback (as our design) paying a randomly selected round is not incentive compatible, because agents have an incentive to experiment.

Regret	1	2	3	4	5	6
Number	550	494	552	617	589	588
Frequency (%)	16.22	14.57	16.28	18.20	17.37	17.35

Table 2: Frequency of the different experienced regret levels in lose situations.

was no feedback during the first choice of each part). In addition, the column in the payoff table which represented the choice of the other player in the previous round was highlighted.

The translated and original instructions including screenshots from the experiment can be found in the Supplementary Materials. After completing the 39 rounds of play, participants answered demographic questions, the Maximization and Regret scale (Schwartz et al., 2002), the Faith in Intuition scale (Epstein et al., 1996; Alós-Ferrer and Hügelschäfer, 2012, 2016), and a 15-item Big-Five questionnaire (Lang et al., 2011, p. 560).[3] The results reported below are robust to the inclusion of those measures as controls, but the measures themselves provided no additional insights.

3 Results

3.1 Classification of the Decision Situations

We are interested in the choices that participants made while seeing the previous round's outcomes, and hence we dropped the first round of each part (where no feedback on a previous decision was possible). For all other decisions, if the participant had achieved the highest possible payoff in the previous round, we classified the following decision as a *win* situation, else as a *lose* situation. The complete data set consists of $144 \times 36 = 5184$ observations, of which 1,794 observations (34.61%) were win situations and 3,390 (65.39%) were lose situations. Regret, defined as the difference between the maximum possible payoff and the realized one, was always 0 in win situations and strictly positive in lose situations. Table 2 contains the frequency of various levels of regret in lose situations, which were not significantly different from a uniform distribution (χ^2-test, $\chi^2_{(5)} = 8.359$, $p = 0.1375$).

The "win-stay, lose-shift" version of reinforcement we focus on prescribed staying with the previously chosen option in win situations and shifting away to another option in lose situations. However, in win situations staying with the previous action is also the prescription of myopic best reply in our games, since if a player assumes that the opponent will not change strategy, repeating the strategy which led to a win is optimal. Further, mere inertia (repeating the previous action no matter what; see, e.g., Alós-Ferrer et al., 2016) also leads to the same prescription. That is, in win situations all three behavioral rules (reinforcement, myopic best reply, and inertia) prescribed to repeat

[3]The German versions of the three scales were taken from Greifeneder and Betsch (2006), Keller et al. (2000), and Gerlitz and Schupp (2005), respectively.

the previous action. There were two possible deviations from this common prescription, i.e. the choice which would have yielded the medium payoff and the one which would have yielded the low payoff (if the opponent stayed put).

In lose situations reinforcement prescribed to shift away from the previous choice to one of the two remaining alternatives. Shifting to the maximum-payoff alternative was aligned with myopic best reply (BR shifts), but shifting to the remaining alternative (which could be a medium- or low-payoff one) was not (non-BR shifts). Since BR shifts just require following the bygone payoff, we focus on a reinforcement rule prescribing BR shifts. By definition, the prescription of inertia was to stay put with the previously-chosen alternative.

The strategy of analysis is as follows. At the first level, we take decisions as a response to the feedback, i.e. we are interested in (Markov) decision rules as studied in the literature on learning in games (e.g., Fudenberg and Levine, 1998). For such an analysis, the origin of the input (feedback) is irrelevant, since one studies the probabilities of responses conditional on the input. Hence we study subject averages (frequency of choices and response times conditional on choice and situation) with 144 observations (one per player) and conduct non-parametric (within) tests. At the second level, we conduct a robustness analysis and reanalyze the data at the block level (since matching was within blocks of four players), creating 36 independent observations for non-parametric tests. That is, at this level each individual observation averages over all decisions of all four players in a block. At the third level, we analyze the data as a panel through the appropriate regression models, controlling for a number of additional factors.

3.2 Choice Data

Figure 1 depicts the average individual choice frequencies conditional on win and lose situations. In win situations, the decision to "stay" (following reinforcement, myopic best reply, and inertia) was significantly more frequent than any other alternative. Stay decisions, with an average individual frequency of 51.57%, where more frequent than shifts to the medium-payoff (26.35%) or the low-payoff action (22.08%). The differences are significant according to Wilcoxon-Signed Rank tests[4] (stay vs. shift to medium, $z = 6.130$, $p < 0.0001$; stay vs. shift to low, $z = 6.615$, $p < 0.0001$). Shifts to the medium-payoff action were also more frequent than shifts to the low-payoff action, and the difference in distributions is also significant ($z = 2.599$, $p = 0.0094$). Tests at the block level ($N = 36$) yielded the same conclusions (WSR tests, stay (50.73%) vs. shift to medium (26.92%), $z = 4.550$, $p < 0.0001$; stay vs. shift to low (22.36%), $z = 4.643$, $p < 0.0001$; shift to medium vs. shift to low, $z = 2.353$, $p = 0.0186$).

In lose situations, shifts to the myopic best reply were also more frequent than other choices. The average frequency of shifts aligned with myopic best reply was 42.74%, compared to 26.41% of non-BR shifts (WSR test, $z = 6.295$, $p < 0.0001$) and 30.84% of

[4]Here and elsewhere, tests are adjusted for multiple comparisons following the Holm-Bonferroni correction.

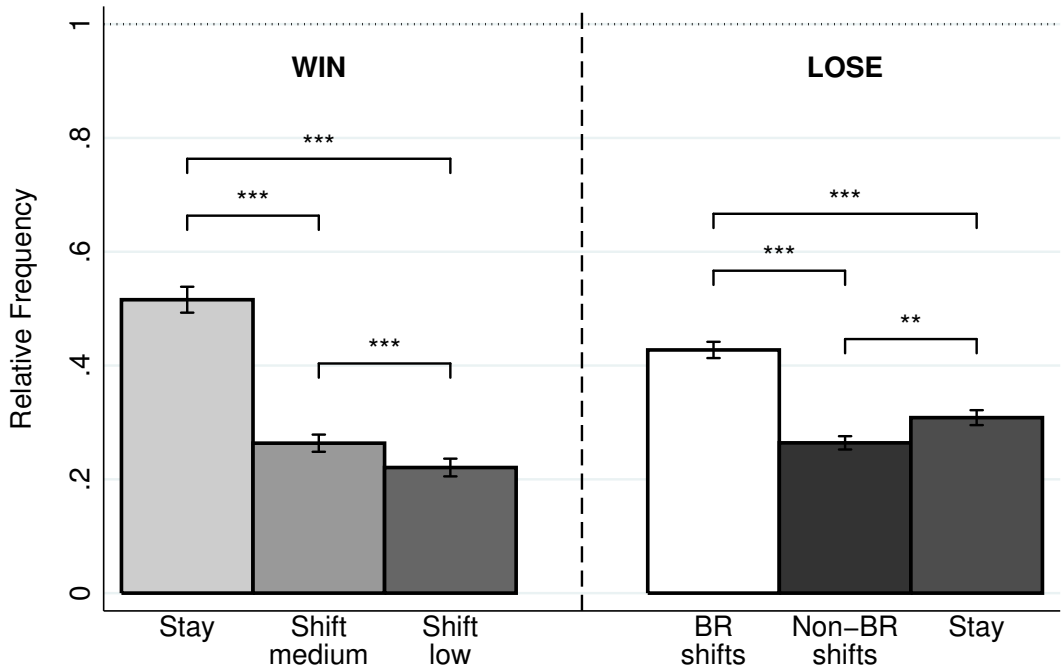

Figure 1: Average individual choice frequencies conditional on Win or Lose situations. Significance levels refer to Wilcoxon Signed-Rank tests, adjusted for multiple comparisons. ** $p < 0.05$, *** $p < 0.01$.

stay choices following inertia (WSR test, $z = 4.443$, $p < 0.0001$). Non-BR shifts were less frequent than inertia decisions (WSR test, $z = -2.108$, $p = 0.0350$). Testing at the block level ($N = 36$) yielded the same conclusions (WSR tests, BR shifts (42.07%) vs. non-BR shifts (26.40%), $z = 4.454$, $p < 0.0001$; BR shifts vs. inertia (31.53%), $z = 3.849$, $p = 0.0002$; non-BR shifts vs. inertia, $z = -2.090$, $p = 0.0367$).

Figure A.1 (left-hand side) shows that the main results above are robust when choice frequencies are analyzed for each game separately. That is, the decision to stay in win situations, respectively to shift to the myopic best reply in lose situations, was more frequent than other choices in each of the three games.

Table 3 presents panel probit regressions with random effects at the individual level and standard errors clustered at the matching block. The independent variable is defined as 1 if the decision followed the prescription of reinforcement / myopic best reply, that is, stay in win situations and shift to the best-payoff action in lose situations. The regressions allow us to control for other variables such as the regret level or learning effects. In all three models, the dummy Win (for win situations) is highly significant and positive, implying that participants were more likely to follow the common prescription of reinforcement and myopic best reply in win situations. That is, win-stay decisions

Reinforcement decision	Model 1	Model 2	Model 3
Win	0.2201***	0.3737***	0.3733***
	(0.0618)	(0.0859)	(0.0849)
Lose × Regret		0.0425**	0.0423**
		(0.0170)	(0.0167)
Normalized Round			−0.1171
			(0.0736)
Part 2 Dummy			0.0301
			(0.0569)
Part 3 Dummy			−0.0323
			(0.0581)
Constant	−0.2075***	−0.3606***	−0.2959***
	(0.0377)	(0.0738)	(0.0887)
LogLikelihood	−3429.1594	−3424.0503	−3421.2967
Wald Test	33.0209***	43.1477***	48.5771***

Table 3: Random-effects panel probit regressions for choices. Independent variable takes value 1 if reinforcement was followed. Standard errors (clustered by 36 matching blocks) in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

are more likely than lose-shift (to the best reply) even though both follow reinforcement and myopic best reply. This asymmetry is as it should be expected from inertia, since a "stay" decision complies with this additional process (Erev and Haruvy, 2016; Alós-Ferrer et al., 2016).[5]

Model 2 adds the interaction Lose × Regret, which captures the effect of the regret level in lose situations (recall that regret is identically zero for win situations). The coefficient is also significantly positive in Models 2 and 3, indicating that being further away from the maximum payoff resulted in a higher probability of following reinforcement. Model 3 controls for learning effects, adding a coefficient for normalized round within a part (rescaled to range from 0 to 1) and individual dummies for each part. None were significant, indicating no evidence of behavioral change over time. That is, consistent with our Markov-rules approach, at the aggregate level there is no evidence that the reliance on reinforcement changed over time. Additional regressions including the personality questionnaires and demographic factors (Table S.1, Supplementary Materials) as controls yielded the same qualitative results, while the controls themselves provided no further insights.

Cameron et al. (2008) remarked that a small number of clusters can lead to over-rejection of standard asymptotic tests and recommending bootstrapping the standard errors. According to their simulations group sizes of 30 already showed a rejection rate close to the intended 5% level. Hence, as a further robustness test, we ran a regression bootstrapping the standard errors with 100 repetitions (Table S.2, Supplementary Materials), which yielded the same conclusions.

[5]This is also confirmed by WSR tests comparing win-stay rates (average 51.57%) and the rate of shifts to the best reply in lose situations (average 42.74%), which are significant both at the individual ($z = 3.275, p = 0.0011$) and the block level ($z = 3.119, p = 0.0018$).

In summary, the regressions indicate that "win-stay" in win situations were comparatively more likely than shifting to the myopically best option in lose situations, but that the "lose-shift" choice was more likely with higher regret. Note that the latter result cannot be explained by a behavioral rule based on myopic best reply (which should be impervious to regret), but is consistent with a general reinforcement process for which the strength of a loss does play a role.

3.3 Cumulative Reinforcement

The behavioral rules we concentrate on rely on information from the most recent period of play only. One can of course ask whether alternative reinforcement rules using information from longer histories can describe the data better. A prominent example is *cumulative proportional reinforcement* (Laslier et al., 2001), which prescribes to choose the action with the highest cumulative payoff over the whole past. Relying only on the simple rules of reinforcement and inertia, which only take the previous round into account, we can account for $3,406$ (65.70%) out of the $5,184$ observations in our data set. In contrast, only $2,214$ observations (42.71%) agree with cumulative proportional reinforcement, and the majority of those observations are also aligned with simple reinforcement and inertia. The combination of inertia and cumulative proportional reinforcement captures $2,813$ observations (54.26%). Only 487 (9.39%) observations are consistent with cumulative proportional reinforcement but not with the other two rules, while $1,080$ (20.83%) are compatible with our simple reinforcement rule but not with the other two rules.

We also ran regressions analogous to Table 3 for cumulative proportional reinforcement (Table S.3, Supplementary Materials). That is, the dependent variable was defined as a dummy taking the value 1 if and only if the decision was consistent with this rule. The regression did not yield any significant results for winning or regret levels, but showed a significantly negative time trend, indicating a lower likelihood of following cumulative proportional reinforcement over time. In conclusion, we view these observations as indicators that it is reasonable to assume Markov decisions rules in this context.

3.4 Response Times

We computed average individual response times conditional on the different situations and choices, dropping the first round of each part. Note that not all participants have response times for each shift or stay, since if for instance a participant never shifted away in win situations, there will be no observations in the corresponding categories. Hence, the tests below will in general have different numbers of observations (tests at the block level, however, always have 36 observations).

Figure 2 depicts the averages of response times for shift and stay decisions, conditional on win or lose situations. The fastest decisions are always those consistent with the common prescription of myopic best reply and reinforcement. Specifically, in win situa-

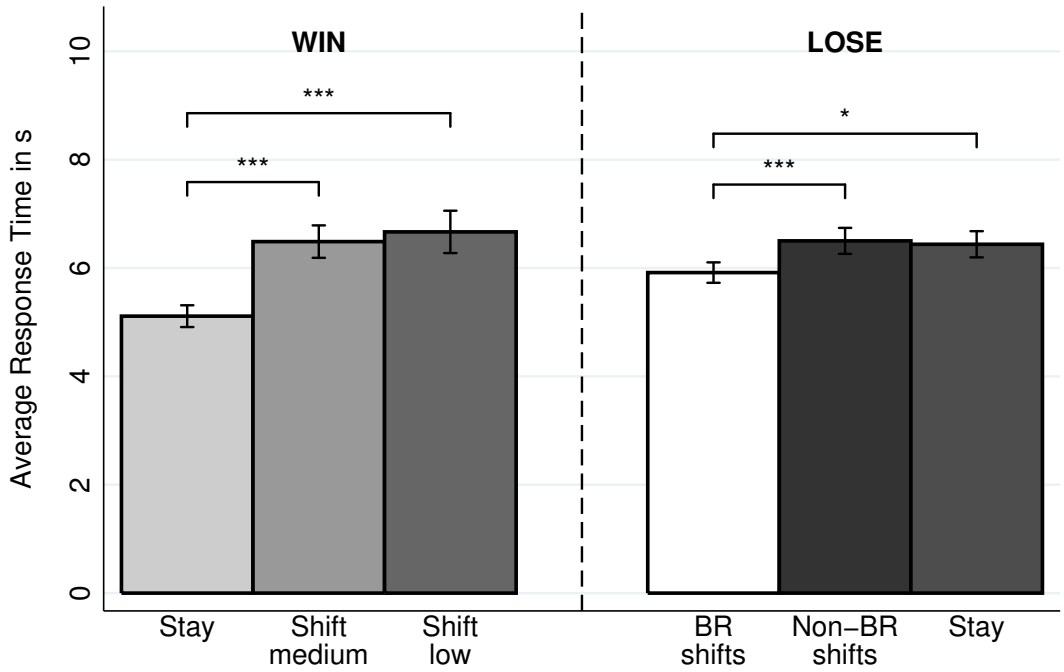

Figure 2: Average individual response times conditional on choice and situation. Significance levels refer to Wilcoxon Signed-Rank tests, adjusted for multiple comparisons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

tions, stay decisions were faster than shifts to the medium-payoff action (average 5.25 s vs. 6.53 s; WSR test, $N = 126$, $z = -4.342$, $p < 0.0001$) and shifts to the low-payoff action (average 5.34 s vs. 6.73 s; WSR test, $N = 112$, $z = -4.121$, $p < 0.0001$). Response times for both kinds of shifts were not significantly different in win situations (shift to medium, 6.34 s; shift to low, 6.53 s; WSR test, $N = 109$, $z = -0.322$, $p = 0.7475$). Testing at the block level ($N = 36$) yields identical conclusions (WSR tests, stay (mean 5.01 s) vs. medium (mean 6.34 s), $z = -3.629$, $p = 0.0006$; stay vs. low (mean 6.48 s), $z = -3.912$, $p = 0.0003$; medium vs. low, $z = -0.471$, $p = 0.6374$).

The same also holds for lose situations. That is, shifts aligned with myopic best reply were faster than non-BR shifts (average 5.91 s vs. 6.50 s; WSR test, $N = 142$, $z = -2.988$, $p = 0.0084$) and stay decisions (average 5.95 s vs. 6.44 s; WSR test, $N = 142$, $z = -2.066$, $p = 0.0777$). Response times for the latter two were not significantly different in lose situations (non-BR shifts, 6.55 s; stay decisions, 6.45 s; WSR test, $N = 140$, $z = 0.774$, $p = 0.4391$). However, tests at the block level ($N = 36$; adjusted, as always, following Holm-Bonferroni) were not significant in this case (WSR tests, BR shifts (5.98 s) vs. non-BR shifts (6.48 s), $z = -1.870$, $p = 0.1231$; BR shifts vs. stay (6.06 s), $z = -0.314$, $p = 0.7534$; non-BR shifts vs. stay, $z = 1.901$, $p = 0.1719$).

Figure A.1 (right-hand side) illustrates the response time analysis above for each game separately. The decision to stay in win situations was faster than other decisions for all Games. The decision to shift to the myopic best reply in lose situations was faster than other shifts and stay decisions for Game 3 and faster than stay decisions for Game 1, but the differences were not significant for Game 2.

Table 4 presents panel regressions for log-transformed response times, with random effects at the individual level and standard errors clustered at the matching block.[6] The analysis confirms that decisions which agree with the prescriptions of reinforcement (and best reply) are significantly faster than other decisions. To see this for win-stay, we look at the dummy "Stay" which indicates whether the decision was to repeat the previous choice or not, i.e. inertia. Its coefficient is highly significant and negative in all models, indicating that (since the interaction Lose $\times$ Stay is included in the models) decisions to stay after a win, hence to stay with the best response, were made significantly faster. This is consistent with the interpretation that many such decisions might be the result of relatively automatic processes.[7] For lose-shift, we look at the dummy "BR-Shift" which captures shifts to the payoff-maximizing option after a lose situation, as in Figure 2 (note that shifting to a best response implies a lose situation). Its coefficient is also significant and negative in all models, again indicating faster decisions.[8]

The regressions also allow us to examine additional questions. This first concerns the difference between win and lose situations. Reinforcement / best-reply decisions were significantly faster in win situations than in lose situations, as revealed by a linear combination test (1) at the bottom of Table 4.[9] The dummy for lose situations is not significant, implying that shifts not following the reinforcement prescription did not differ in response times across win and lose situations.

The second additional observation concerns inertia compared to shifts. All models include an interaction term for stay decisions in lose situations (Lose $\times$ Stay), that is, for following inertia. Concerning inertia and non-BR shifts, the linear combination test (2) in Table 4 is only marginally significant, and only in model 3. That is, there is

[6]Response times are nonnegative and their distribution is strongly right-skewed, while the distribution of log-transformed response times typically shows a normally-distributed shape (e.g., Fischbacher et al., 2013). Using Shapiro-Wilk W tests at the individual level, the hypothesis of normality was only rejected for 19 of the 144 subjects at the 5% level, and there were only 7 cases with significance between 5% and 10%. In contrast, using regular response times, the hypothesis of normality is rejected in 113 cases at the 5% level, and in 11 further cases at the 10% level.

[7]Some of the non-stay, slower decisions might be the result of more complex decision rules, as e.g. best-responding to a predicted best response.

[8]The omitted category are (non-BR) shifts after a win. Since the non-parametric analysis did not find significant differences between shifts to medium and low-payoff actions, we do not distinguish them further. An additional regression including a dummy for shifts to the medium-payoff action in win situations (Table S.4, Supplementary Materials) yields the same conclusions, with the additional coefficient not being significantly different from 0.

[9]This difference cannot be explained through pure best-reply behavior, since best reply only depends on the opponent's previous choice and not on whether there was a previous win or loss. Reinforcement yields prescriptions conditional on the situation (win-stay, lose-shift) and could hence account for such difference. Alternatively, one could argue that, for unmodeled reasons, some best-reply decisions (e.g. win-stay) are less cognitively demanding than others.

ln(response time)	Model 1	Model 2	Model 3
Stay	-0.2753^{***}	-0.2752^{***}	-0.2912^{***}
	(0.0412)	(0.0412)	(0.0411)
Lose	0.0086	0.0037	-0.0039
	(0.0366)	(0.0421)	(0.0424)
Lose × Stay	0.2202^{***}	0.2205^{***}	0.2233^{***}
	(0.0491)	(0.0487)	(0.0476)
BR-Shift	-0.0633^{**}	-0.0634^{**}	-0.0680^{**}
	(0.0321)	(0.0322)	(0.0313)
Lose x Regret		0.0013	0.0024
		(0.0074)	(0.0077)
Normalized Round			-0.3446^{***}
			(0.0407)
Part 2 Dummy			-0.0790^{*}
			(0.0409)
Part 3 Dummy			0.0224
			(0.0503)
Constant	1.6494^{***}	1.6494^{***}	1.8643^{***}
	(0.0440)	(0.0441)	(0.0537)
R^2	0.0256	0.0256	0.0555
Wald Test	105.5638^{***}	111.5129^{***}	298.8777^{***}
Linear Combination Tests:			
(1) Lose + BR-Shift vs. Stay	0.2205^{***}	0.2155^{***}	0.2192^{***}
	(0.0292)	(0.0391)	(0.0414)
(2) Stay + Lose × Stay	-0.0551	-0.0547	-0.0678^{*}
	(0.0410)	(0.0404)	(0.0384)
(3) BR-Shift vs. Stay + Lose × Stay	-0.0082	-0.0087	-0.0002
	(0.0286)	(0.0277)	(0.0270)

Table 4: Panel regressions for log-transformed response times. Standard errors (clustered by 36 matching blocks) in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

only weak evidence that inertia decisions might be faster than non-BR shifts in lose situations. Concerning inertia and BR shifts, however, the linear combination test (3) comparing BR-shift decisions with stay decisions in lose situations was not significant, in contrast with the non-parametric results.

Models 2 and 3 also add the interaction with the regret level (in lose situations, as in win situations there is no regret), which is not significant, implying that the level of regret experienced did not affect the response time. This is not at odds with our previous finding that higher regret levels induce more reinforcement, because the BR-Shift coefficient already captures the faster response times in lose situations. Model 3 controls for learning effects as in the regression on choice data. As standard in choice experiments, participants became faster as their familiarity with the interface and the situation increased, as indicated by a significantly negative coefficient for Normalized Round. Participants also became faster after the first part (although the effect is only weakly significant), but there was no improvement of response times in the third part

compared to the first. Additional regressions including questionnaires and demographics (Table S.5, Supplementary Materials) as controls yielded the same qualitative results. Further, an additional regression bootstrapping the standard errors with 100 repetitions (Table S.6, Supplementary Materials) yielded the same conclusions.

4 Discussion

In our experiment, we examined a behavioral rule based on the simplest reinforcement principle, "win-stay, lose-shift," as a possible cognitive short-cut for myopic best reply. After a previous win, the rule prescribes to repeat the previous choice. After a previous lose situation, the rule prescribes to shift away from it, and available information allows to focus on the specific shift favored by myopic best reply. Reinforcement processes are known to be automatic and associated with short response times, while myopic best reply, which involves explicit maximization, should be a more deliberative, slower process.

In win situations, win-stay was a strong driver of behavior. Such decisions occurred more often and faster than other actions (results confirmed by non-parametric tests both at the individual and block level and by panel regressions). Choice data alone could be justified by either reliance on an automatic, impulsive reinforcement process or a more cognitive, deliberative explicit maximization. However, win-stay decisions were the fastest decisions observed in our experiment, indicating that a more automatic process was at work compared to slower decisions. This leads to the interpretation that stay responses, even though they correspond to myopic bet response, most likely often followed a more automatic process.

In lose situations, shift rates (to the best reply) were sensitive to the magnitude of experienced regret, in agreement with reinforcement learning but in contrast to myopic best reply, which would predict the best-reply rates to be independent from the cardinal difference between experienced and maximum possible payoffs.[10] Also, shifts to the myopic best reply were more frequent and faster than other choices, confirming the general interpretation of a more automatic underlying process, although response-time evidence was less strong than in the case of win situations.

Regression results on response times suggest that stay decisions (consistent with inertia) might actually be as fast as choices following reinforcement or best reply, in agreement with evidence on inertia as an additional, automatic process (Alós-Ferrer et al., 2016). This also agrees with the observation that, in lose situations, stay decisions occurred more often than shifts which did not follow the maximum payoff. However, shifts to the myopic best reply were more frequent with higher regret, an effect incompatible with a pure best-reply interpretation, but perfectly aligned with general reinforcement processes in which the strength of the loss does play a role.

[10] In this setting, reinforcement could be seen as a gradual, payoff-dependent version of best reply, giving rise to behavior in line with quantal response models (McKelvey and Palfrey, 1995).

Our experiment goes beyond previous paradigms on win-stay, lose-shift rules (which have typically employed binary-choice settings) and shows evidence on the relevance of simple reinforcement processes in strategic settings. The evidence can and should be seen as a general caveat, in the sense that evidence in favor of myopic best reply in games might often be confounded with the workings of reinforcement, the basic building block of human learning (Sutton and Barto, 1998; Daw and Tobler, 2014; Achtziger et al., 2015). In this and other cases of overlapping behavioral rules, the analysis of response times might provide valuable evidence.

Appendix

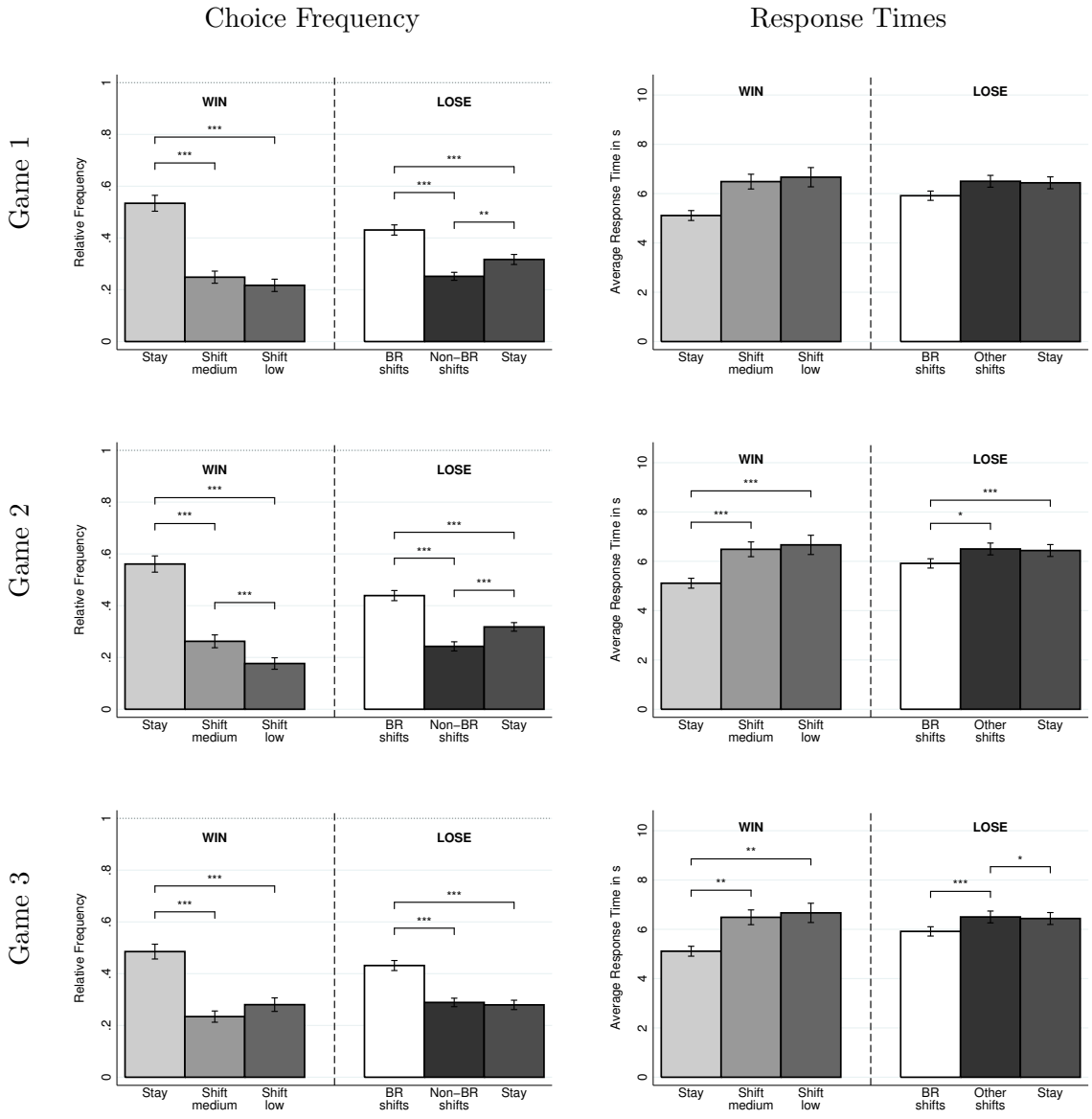

Figure A.1: Average individual choice frequencies and response times conditional on choice and situation. Significance levels refer to Wilcoxon Signed-Rank tests, adjusted for multiple comparisons. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.jebo.2018.06.014.

References

Achtziger, A. and Alós-Ferrer, C. (2014). Fast or Rational? A Response-Times Study of Bayesian Updating. *Management Science*, 60(4):923–938.

Achtziger, A., Alós-Ferrer, C., Hügelschäfer, S., and Steinhauser, M. (2015). Higher Incentives Can Impair Performance: Neural Evidence on Reinforcement and Rationality. *Social Cognitive and Affective Neuroscience*, 10(11):1477–1483.

Alós-Ferrer, C. and Hügelschäfer, S. (2012). Faith in Intuition and Behavioral Biases. *Journal of Economic Behavior and Organization*, 84(1):182–192.

Alós-Ferrer, C. and Hügelschäfer, S. (2016). Faith in Intuition and Cognitive Reflection. *Journal of Behavioral and Experimental Economics*, 64:61–70.

Alós-Ferrer, C., Hügelschäfer, S., and Li, J. (2016). Inertia and Decision Making. *Frontiers in Psychology*, 7 (169):1–9.

Alós-Ferrer, C., Hügelschäfer, S., and Li, J. (2017). Framing Effects and the Reinforcement Heuristic. *Economics Letters*, 156:32–35.

Alós-Ferrer, C. and Strack, F. (2014). From Dual Processes to Multiple Selves: Implications for Economic Behavior. *Journal of Economic Psychology*, 41:1–11.

Alós-Ferrer, C. and Weidenholzer, S. (2007). Partial Bandwagon Effects and Local Interactions. *Games and Economic Behavior*, 61:1–19.

Alós-Ferrer, C. and Weidenholzer, S. (2008). Contagion and Efficiency. *Journal of Economic Theory*, 143:251–274.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in Experiments: A Theoretical Analysis. *Journal of Political Economy*, forthcoming.

Baron, J. and Hershey, J. C. (1988). Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology*, 54(4):569–579.

Börgers, T. and Sarin, R. (1997). Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory*, 77(1):1–14.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, 90(3):414–427.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental Methods: Pay One or Pay All. *Journal of Economic Behavior and Organization*, 131:141–150.

Charness, G. and Levin, D. (2005). When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect. *American Economic Review*, 95(4):1300–1309.

Daw, N. D. (2012). Model-Based Reinforcement Learning as Cognitive Search: Neurocomputational Theories. In Todd, P. M., Hills, T. T., and Robbins, T. W., editors, *Cognitive Search: Evolution, Algorithms and the Brain*, pages 195–208. MIT Press, Cambridge, MA.

Daw, N. D. and Tobler, P. (2014). Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. In Glimcher, P. W. and Fehr, E., editors, *Neuroeconomics: Decision Making and the Brain*, pages 283–298. Academic Press, London, 2nd edition.

Dillon, R. L. and Tinsley, C. H. (2008). How Near-Misses Influence Decision Making Under Risk: A Missed Opportunity for Learning. *Management Science*, 54(8):1425–1440.

Epstein, S., Pacini, R., Denes-Raj, V., and Heier, H. (1996). Individual Differences in Intuitive-Experiential And Analytical-Rational Thinking Styles. *Journal of Personality and Social Psychology*, 71(2):390–405.

Erev, I. and Haruvy, E. (2016). Learning and the Economics of Small Decisions. In Kagel, J. and Roth, A. E., editors, *Handbook of Experimental Economics, Volume 2*, pages 638–716. Princeton University Press.

Erev, I. and Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88(5):848–881.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics*, 10(2):171–178.

Fischbacher, U., Hertwig, R., and Bruhin, A. (2013). How to Model Heterogeneity in Costly Punishment: Insights from Responders' Response Times. *Journal of Behavioral Decision Making*, 26(5):462–476.

Fudenberg, D. and Levine, D. K. (1998). *The Theory of Learning in Games*. The MIT Press, Cambridge, Massachusetts.

Gerlitz, J.-Y. and Schupp, J. (2005). Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP. *DIW Research Notes 2005-04*.

Gershman, S. J., Pesaran, B., and Daw, N. D. (2009). Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *The Journal of Neuroscience*, 29(43):13524–13531.

Gittins, J. C. (1979). Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society, Series B*, 41:148–177.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. Wiley, New York.

Greifeneder, R. and Betsch, C. (2006). Lieber die Taube auf dem Dach! *Zeitschrift für Sozialpsychologie*, 37(4):233–243.

Greiner, B. (2015). Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. *Journal of the Economic Science Association*, 1:114–125.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15:534–539.

Holroyd, C. B. and Coles, M. G. (2002). The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity. *Psychological Review*, 109:679–709.

Hügelschäfer, S. and Achtziger, A. (2017). Reinforcement, Rationality, and Intentions: How Robust is Automatic Reinforcement Learning in Economic Decision Making? *Journal of Behavioral Decision Making*, 30(4):913–932.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, 93(5):1449–1475.

Kandori, M. and Rob, R. (1995). Evolution of Equilibria in the Long Run: A General Theory and Applications. *Journal of Economic Theory*, 65:383–414.

Keller, J., Bohner, G., and Erb, H.-P. (2000). Intuitive und heuristische Urteilsbildung — verschiedene Prozesse? Präsentation einer deutschen Fassung des 'Rational–Experiential Inventory' sowie neuer Selbstberichtskalen zur Heuristiknutzung. *Zeitschrift für Sozialpsychologie*, 31(2):87–101.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., and Wagner, G. G. (2011). Short Assessment of the Big Five: Robust Across Survey Methods Except Telephone Interviewing. *Behavior Research Methods*, 43(2):548–567.

Laslier, J.-F., Topol, R., and Walliser, B. (2001). A Behavioral Learning Process in Games. *Games and Economic Behavior*, 37:340–366.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38.

Schönberg, T., Daw, N. D., Joel, D., and O'Doherty, J. P. (2007). Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners During Reward-Based Decision Making. *The Journal of Neuroscience*, 27(47):12860–12867.

Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology*, 80:1–27.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., and Lehman, D. R. (2002). Maximizing Versus Satisficing: Happiness is a Matter of Choice. *Journal of Personality and Social Psychology*, 83(5):1178.

Strack, F. and Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review*, 8(3):220–247.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. MacMillan (see also Hafner Publishing Co., 1970), NY.

Vega-Redondo, F. (1997). The Evolution of Walrasian Behavior. *Econometrica*, 65(2):375–384.

Weibull, J. (1995). *Evolutionary Game Theory*. The MIT Press, Cambridge, Massachusetts.

Wimmer, G. E., Daw, N. D., and Shohamy, D. (2012). Generalization of Value in Reinforcement Learning by Humans. *European Journal of Neuroscience*, 35(7):1092–1104.